

OPTIMASI PARAMETER K PADA ALGORITMA *K-NEAREST NEIGHBOUR* UNTUK KLASIFIKASI PENYAKIT DIABETES MELLITUS

Indrayanti¹, Devi Sugianti¹, M. Adib Al Karomi^{1*}

STMIK Widya Pratama Pekalongan

*E-mail: adib.comp@gmail.com

Abstrak

Diabetes Mellitus merupakan salah satu penyakit kronis yang mematikan. Penyakit yang juga dikenal dengan nama penyakit kencing manis ini terjadi akibat kadar glukosa di dalam darah terlalu tinggi. Diabetes Mellitus banyak diteliti di banyak negara pada saat ini karena peningkatan penderita yang banyak dan sangat mengkhawatirkan. Menurut WHO saat ini lebih dari 246 juta jiwa menderita diabetes dan diperkirakan akan meningkat menjadi 380 juta jiwa pada tahun 2025 apabila tidak dilakukan penanganan yang serius. Diabetes menyebabkan penyakit lain / komplikasi yang setiap tahunnya mengakibatkan kematian hingga 3,8 juta jiwa. Data mining merupakan kegiatan menemukan sebuah pola, aturan dan pengetahuan baru dari sebuah dataset. Salah satu fungsi mayor data mining adalah klasifikasi. KNN merupakan salah satu algoritma klasifikasi data mining terbaik dan banyak digunakan. Algoritma KNN bekerja dengan cara menghitung kedekatan data testing dengan keseluruhan data training. K dalam KNN merupakan variabel jumlah tetangga terdekat yang akan diambil untuk proses klasifikasi. Jumlah K=1 akan membuat hasil klasifikasi terasa kalu karena hanya memperhitungkan satu tetangga terdekat atau satu record karakteristik data terdekat. Sedangkan jumlah K yang terlalu banyak akan menghasilkan klasifikasi yang samar. Penelitian ini menghasilkan K terbaik pada percobaan K=13 dengan akurasi 75,14%. K=13 merupakan nilai k paling optimal diantara percobaan klasifikasi KNN menggunakan nilai K=1 sampai dengan K=49.

Kata Kunci : Kencing manis, Peningkatan akurasi KNN

1. PENDAHULUAN

1.1.Latar Belakang Masalah

Diabetes Mellitus atau biasa disebut juga dengan penyakit kencing manis merupakan penyakit yang terjadi akibat kadar glukosa di dalam darah terlalu tinggi. Kadar gula didalam darah dapat meningkat salah satunya disebabkan karena tubuh tidak dapat melepaskan atau menggunakan insulin secara normal. Setiap individu memiliki kadar glukosa yang bervariasi, kadar glukosa ini akan meningkat setelah makan kemudian akan kembali normal dalam waktu dua jam (Barakat et al. 2010). Pada umumnya kadar glukosa darah akan meningkat secara ringan pada usia muda, tetapi kadar glukosa ini akan mengalami peningkatan yang progresif pada usia lebih dari 50 tahun. Peningkatan kadar glukosa ini akan lebih terasa pada orang dengan gaya hidup pasif atau jarang beraktifitas. Insulin merupakan hormon yang dilepaskan oleh *pancreas* dan merupakan zat utama yang bertanggungjawab dalam mempertahankan kadar glukosa darah yang tepat. Insulin bertugas memindahkan glukosa ke dalam sel sehingga dapat menghasilkan energi. *Diabetes Mellitus* dapat terjadi ketika tubuh tidak menghasilkan insulin yang cukup untuk mempertahankan glukosa darah normal atau jika sel tidak memberikan respon yang tepat terhadap insulin.

Diabetes Mellitus merupakan salah satu penyakit kronis yang mematikan. Penyakit ini juga merupakan jenis penyakit yang banyak diamati di banyak negara saat ini. Penyakit ini terus dan menjadi semakin meningkat pada tingkat yang sangat mengkhawatirkan (Temurtas, Yumusak, and Temurtas 2009). Menurut Report WHO (Report of a WHO / IDF Consultation 2006) saat ini ada 246 juta penderita diabetes diseluruh dunia, dan jumlah ini diperkirakan akan meningkat menjadi 380 juta pada tahun 2025. Diabetes menyebabkan penyakit lain/komplikasi yang setiap tahunnya mengakibatkan kematian 3,8 juta jiwa. Komplikasi yang lebih sering terjadi dan mematikan akibat diabetes adalah serangan jantung dan stroke. Sebagian besar kematian terjadi karena kadar glukosa mengalami kenaikan terus menerus sehingga berakibat rusaknya pembuluh darah, saraf dan srtuktur internal lainnya. Indonesia menempati urutan ke enam di dunia sebagai Negara dengan jumlah penderita *Diabetes Mellitus* terbanyak setelah India, Cina, Unisoviet,

Jepang dan Brasil. Pada tahun 2006 jumlah penderita *Diabetes Mellitus* di Indonesia mencapai 14 juta jiwa, jika peningkatan penderita *Diabetes Mellitus* meningkat 230.000 jiwa setiap tahunnya, maka bisa kita bayangkan berapa banyak jumlah penderita *Diabetes Mellitus* pada tahun 2017.

Dalam bidang kedokteran terdapat banyak catatan penderita penyakit salah satunya data penyakit diabetes. Data yang sangat banyak belum dapat digunakan apabila tidak ada informasi atau kesimpulan dari data tersebut. Bahkan data yang banyak justru dapat menjadi sampah dan tidak berguna. Suatu proses ekstraksi untuk mencari informasi dalam data yang belum diketahui sebelumnya dikenal dengan istilah data mining (Witten, Frank, and Hall 2011). Data mining menggunakan teknik pengenalan pola seperti statistik dan matematika untuk menemukan pola dari data atau kasus lama (Larose 2005). Data mining merupakan kegiatan yang meliputi pengumpulan dan pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data (Santosa 2007). Data mining merupakan ilmu yang memanfaatkan data yang sebelumnya kurang terpakai untuk mendapatkan suatu informasi atau pengetahuan baru.

Salah satu fungsi utama data mining adalah klasifikasi. Klasifikasi banyak digunakan untuk menentukan keputusan sesuai pengetahuan baru yang didapat dari pengolahan data lampau menggunakan perhitungan suatu algoritma. Teknik klasifikasi dapat diterapkan dalam semua bidang misalnya dalam bidang kesehatan (Christobel and Sivaprakasam 2011), bidang pendidikan (Ragab et al. 2014), bidang teknik bangunan (Ashari, Paryudi, and Tjoa 2013), bidang jaringan komputer serta banyak digunakan dalam bidang lain. Dalam dataset klasifikasi terdapat satu atribut tujuan atau dapat pula disebut dengan atribut label. Atribut inilah yang akan dicari dari data baru dengan dasar atribut lain pada data lampau. Banyaknya atribut dapat mempengaruhi performa suatu algoritma (Prasetyo 2012). Masalah klasifikasi pada dasarnya adalah sebagai berikut (Susanto and Suryadi 2010):

1. Masalah Klasifikasi berangkat dari data *training* yang tersedia.
2. Data *training* akan diolah dengan menggunakan algoritma klasifikasi.
3. Masalah klasifikasi berakhir dengan dihasilkannya sebuah pengetahuan yang direpresentasikan dalam bentuk diagram, aturan atau pengetahuan.

Beberapa algoritma dapat digunakan untuk melakukan tugas klasifikasi (Han and Kamber 2006). Salah satu algoritma klasifikasi data mining yang terbaik adalah *K-Nearest Neighbour (KNN)* (Wu et al. 2007). Salah satu kelemahan dari algoritma KNN adalah dalam penentuan variabel K (Al Karomi 2015). Nilai K yang terlalu besar akan membuat hasil klasifikasi semakin kabur. Sedangkan apabila nilai K yang digunakan adalah 1 akan mengakibatkan hasil klasifikasi terasa kaku. Penelitian ini menghitung nilai K paling optimal algoritma K-NN untuk Klasifikasi Penyakit *Diabetes Mellitus*.

2. METODE PENELITIAN

Metode penelitian yang digunakan dalam penelitian ini adalah metode eksperimental. Tahapan dalam metode eksperimental yang dilakukan adalah sebagai berikut:

2.1. Metode Pengumpulan Data

Tahapan pertama yang dilakukan dalam penelitian ini adalah pengumpulan data. Data yang akan digunakan dalam penelitian ini adalah data *public* yaitu data *Pima Indian Diabetes Data (PIDD)* dari *Uci Machine Learning Repository*. Dataset ini banyak digunakan oleh peneliti untuk menguji algoritma klasifikasi. Dataset ini berisikan 768 *record* dengan 9 atribut yang salah satunya adalah atribut tujuan atau atribut label. Tabel 3.1 merupakan atribut dataset beserta deskripsinya secara lebih jelas. Dataset ini merupakan dataset yang banyak digunakan untuk klasifikasi penyakit diabetes dan dapat diunduh di url: <https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes>.

Tabel 1. Atribut Dataset beserta deskripsinya

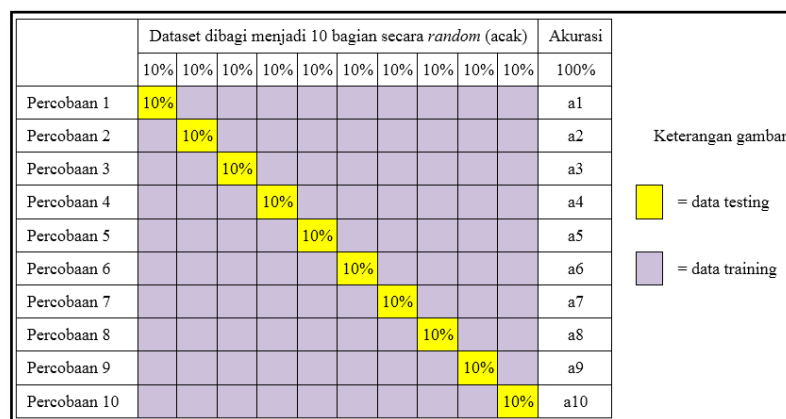
Atribut	Singkatan	Deskripsi	Satuan	Tipe Data
Pregnant	Pregnant	Banyaknya kehamilan	-	Numerik
Plasma-Glucose	Glucose	Kadar glukosa dua jam setelah makan	Mg/dL	Numerik
Diastolic Blood-Pressure	DBP	Tekanan darah	Mm Hg	Numerik
Triceps Skin Fold Thickness	TSFT	Ketebalan kulit	mm	Numerik
Insulin	INS	Insulin	mu U/ml	Numerik
Body Mass Index	BMI	Berat Tubuh	Kg/m2	Numerik
Diabetes pedigree function	DPF	Riwayat Keturunan yang terkena diabetes	-	Numerik
Age	Age	Umur	Years	Numerik
Class variable	Class	Positif diabetes (1) dan negatif diabetes (0)	-	Nominal

2.2. Desain Eksperimen dan Pengujian Algoritma

Dalam tahapan desain dan eksperimen algoritma dilakukan dengan menggunakan aplikasi bantu yaitu rapid miner. Aplikasi rapid miner sengaja digunakan karena dapat digunakan dalam berbagai platform seperti windows dan linux. Selain itu aplikasi ini juga gratis dan tersedia pembaruan setiap bulannya. Dalam penggunaannya aplikasi ini mudah digunakan. Pengguna cukup mempersiapkan dataset kemudian diaplikasikan dengan cara *drag and drop* pada aplikasi untuk mendesain dan melakukan perhitungan. Beberapa algoritma populer dan terbaik juga telah tersedia dalam aplikasi ini. Secara keseluruhan tahapan penelitian ini terbagi menjadi beberapa proses antara lain:

2.2.1. Validasi

Validasi merupakan proses pengujian performa algoritma. Pada umumnya validasi dilakukan dengan mengulang proses perhitungan sampai beberapa kali. Proses validasi dalam penelitian ini menggunakan *cross validation*. *Cross validation* adalah membagi dataset menjadi dua bagian dengan satu bagian dijadikan data *training* dan bagian yang lain dijadikan data *testing*. Beberapa penelitian membagi data menjadi 10 bagian, 90% dijadikan *training* dan 10 lainnya digunakan sebagai *testing*. Proses ini dilakukan berulang sampai dengan 10 kali hingga semua *record* data mendapatkan bagian menjadi data *testing*. Proses ini dikenal juga dengan istilah *10 folds cross validation*. *10 folds cross validation* banyak digunakan peneliti karena terbukti menghasilkan performa algoritma yang lebih stabil. Gambar 1 merupakan representasi dari *10 folds cross validation*.

**Gambar 1. Representasi 10 folds cross validation**

2.2.2. Pengukuran akurasi algoritma

Pengukuran akurasi merupakan tahapan untuk membuktikan tingkat performa suatu algoritma terhadap dataset yang digunakan. Dalam penelitian ini digunakan *confusion matrix* sebagai alat ukur performa algoritma klasifikasi. *Confusion matrix* atau matrik kebingungan merupakan sebuah perhitungan yang membandingkan dataset dengan hasil klasifikasi sesuai dengan data sebenarnya dengan jumlah keseluruhan data. Hasil akhir dari matrik ini adalah tingkat akurasi dengan satuan persen (%). Tingkat akurasi ini yang nantinya dijadikan acuan para peneliti terkait performa algoritma klasifikasi tersebut. *Confusion Matrix* adalah evaluasi dari sebuah klasifikasi data mining yang direpresentasikan menjadi tabel (Gorunescu 2011). *Confusion matrix* berisi informasi perbandingan label hasil klasifikasi dengan label sebenarnya. Tabel 2 menggambarkan *confusion matrix* dengan dua label yaitu yes dan no.

Tabel 2. Confusion Matrix (Gorunescu 2011)

Classification		Predicted class	
		Class: YES	Class: NO
Observed class	Class YES	<i>a</i> True Positive (TP)	<i>b</i> False Negative (FN)
	Class NO	<i>c</i> False Positive (FP)	<i>d</i> True Negative (TN)

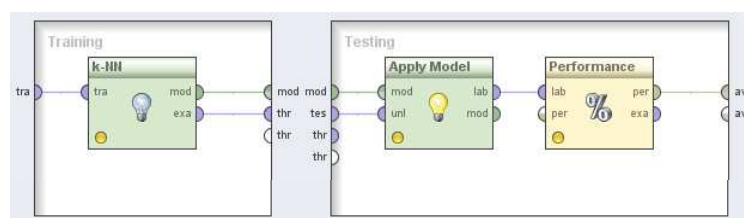
Dari tabel 1 dapat dihitung tingkat akurasi dari sebuah model algoritma dengan menggunakan persamaan sebagai berikut:

$$\text{Accuracy} = \frac{a + d}{a + b + c + d} = \frac{TP + TN}{TP + TN + FP + FN}$$

Keterangan:

- a : hasil klasifikasi positif dengan klas sebenarnya positif
- b : hasil klasifikasi negatif dengan klas sebenarnya positif
- c : hasil klasifikasi positif dengan klas sebenarnya negatif
- d : hasil klasifikasi negatif dengan klas sebenarnya negatif

Dalam aplikasi rapid miner proses perhitungan performa algoritma dapat dijelaskan sesuai gambar 2 berikut. Dalam data training digunakan algoritma knn dan dalam testing digunakan apply model serta prformance untuk menghasilkan matrik kebingungan atau *confusion matrix*. Hasil matrik dari aplikasi dapat dilihat pada gambar 3.



Gambar 2. Confusion matrix model dalam aplikasi rapid miner

accuracy: 74.48% +/- 3.83% (mikro: 74.48%)			
	true tested_positive	true tested_negative	class precision
pred. tested_positive	150	78	65.79%
pred. tested_negative	118	422	78.15%
class recall	55.97%	84.40%	

Gambar 3. Hasil akurasi dari confusion matrix

2.3. Evaluasi Hasil dan Tingkat Akurasi

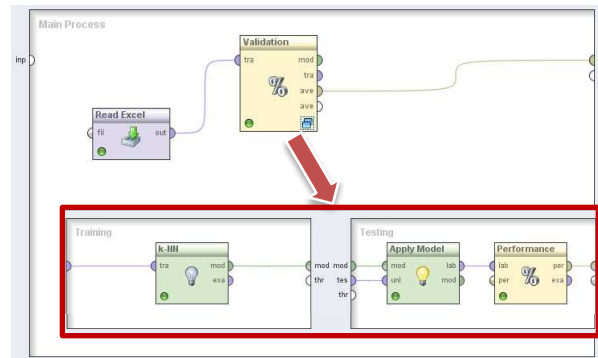
Tahap evaluasi merupakan tahap akhir dari penelitian ini. Dalam tahap ini akan dibandingkan beberapa hasil tingkat akurasi dari algoritma KNN dengan nilai k=1 sampai

dengan nilai $k=49$. Hasil ini akan dicatat lalu dibandingkan satu sama lain untuk menetapkan nilai k yang paling optimal yaitu penggunaan KNN dengan tingkat akurasi tertinggi.

3. HASIL DAN PEMBAHASAN

3.1.1. Hasil

Penelitian ini menggunakan rapid miner dengan percobaan menggunakan nilai K mulai dari 1 sampai dengan 49. Gambar 4 merupakan susunan dataset dan algoritma yang dilakukan dalam penelitian ini.



Gambar 4. Tampilan pada program rapid miner

Dari percobaan perhitungan klasifikasi penyakit diabetes mellitus menggunakan rapid miner, akurasi KNN dengan nilai k yang digunakan dapat dilihat pada tabel 3 berikut:

Tabel 3. Akurasi knn dari berbagai nilai k

k	1	3	5	7	9	11	13	15	17
akurasi	68.24	70.84	72.01	73.06	73.84	74.48	75.14	75	74.87

k	19	21	23	25	27	29	31	33	35
akurasi	73.83	73.31	74.48	74.48	74.61	74.6	74.35	73.05	73.44

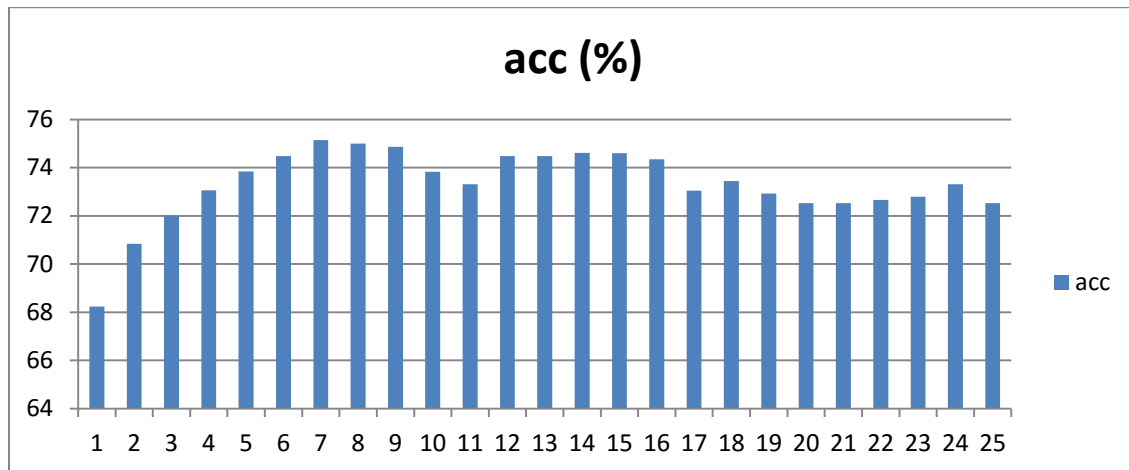
k	37	39	41	43	45	47	49		
akurasi	72.92	72.53	72.53	72.66	72.79	73.31	72.53		

3.2. Pembahasan

Hasil penelitian menunjukkan bahwa nilai $K=13$ merupakan nilai k yang paling optimal dengan tingkat akurasi sebesar 75,14%. Hasil prosentase ini didapatkan dari proses percobaan sebanyak 10 kali dengan data yang diacak (10 fold cross validation) kemudian hasil klasifikasi yang muncul dibandingkan dengan data sebenarnya. Perhitungan ini menggunakan *confussion matrix* untuk menentukan prosentase data yang sesuai dengan kenyataan dibandingkan jumlah keseluruhan data yang ada. Nilai k yang digunakan dalam penelitian ini seluruhnya adalah nilai ganjil dikarenakan label atau hasil akhir dari klasifikasi hanya memiliki 2 kemungkinan yaitu positif dan negatif. Penentuan hasil klasifikasi untuk nilai k lebih dari 1 digunakan metode hasil terbanyak atau mayoritas hasil klasifikasi.

4. KESIMPULAN

Berdasarkan hasil penelitian yang dilakukan terhadap dataset penyakit diabetes. Algoritma klasifikasi KNN dengan nilai $k=13$ menempati peringkat akurasi tertinggi yaitu 75,14%. Grafik kenaikan dan penurunan tingkat akurasi algoritma knn untuk klasifikasi penyakit diabetes dapat dilihat pada gambar 4 berikut.



Gambar 4. Grafik tingkat akurasi klasifikasi penyakit diabetes algoritma knn

DAFTAR PUSTAKA

- Ashari, Ahmad, Iman Paryudi, and A Min Tjoa. 2013. "Performance Comparison between Naïve Bayes , Decision Tree and K-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool" 4 (11): 33–39.
- Barakat, Nahla H, Andrew P Bradley, Senior Member, and Mohamed Nabil H Barakat. 2010. "Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus." *IEEE Transaction on Information Technology in Biomedicine* 14 (4): 1114–20.
- Christobel, Angeline, and D.r Sivaprakasam. 2011. "An Empirical Comparison of Data Mining Classification Methods" 3 (2): 24–28.
- Gorunescu, Florin. 2011. *Data Mining: Concept, Models and Techniques*. Vol 12. Berlin: Heidelberg: Springer Berlin Heidelberg.
- Han, Jiawei, and Micheline Kamber. 2006. *Data Mining: Concepts and Techniques Second Edition*. Elsevier. Elsevier.
- Karomi, M Adib Al. 2015. "Optimasi Parameter K Pada Algoritma KNN Untuk Klasifikasi Heregistrasi Mahasiswa Program Studi Teknik Informatika STMIK Widya Pratama Jl . Patriot 25 Pekalongan Email : Adib.comp@gmail.com." *IC-TECH X* (285): 5.
- Larose, Daniel T. 2005. *Discovering Knowledge in Data: An Introduction to Data Mining*. John Wiley & Sons.
- Prasetyo, Eko. 2012. *Data Mining Konsep Dan Aplikasi Menggunakan Matlab*. Yogyakarta: Andi Offset.
- Ragab, Abdul Hamid M., Amin Y. Noaman, Abdullah S. Al-Ghamdi, and Ayman I. Madbouly. 2014. "A Comparative Analysis of Classification Algorithms for Students College Enrollment Approval Using Data Mining." *Proceedings of the 2014 Workshop on Interaction Design in Educational Environments - IDEE '14*. New York, New York, USA: ACM Press, 106–13. doi:10.1145/2643604.2643631.
- Report of a WHO / IDF Consultation. 2006. "Definition and Diagnosis of Diabetes Mellitus and Intermediate Hyperglycemia." *WHO Library Cataloguing-in-Publication Data*. 1211 Geneva 27, Switzerland.
- Santosa, Budi. 2007. *Data Mining Teknik Pemanfaatan Data Untuk Keperluan Bisnis*. Edisi Pert. Yogyakarta: Graha Ilmu.
- Susanto, Sani, and Dedi Suryadi. 2010. *Pengantar Data Mining: Menggali Pengetahuan Dari Bongkahan Data*. Yogyakarta: Andi Offset.
- Temurtas, Hasan, Nejat Yumusak, and Feyzullah Temurtas. 2009. "A Comparative Study on Diabetes Disease Diagnosis Using Neural Networks." *Expert Systems with Applications* 36 (4). Elsevier Ltd: 8610–15. doi:10.1016/j.eswa.2008.10.032.
- Witten, Ian H, Eibe Frank, and Mark A. Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques 3rd Edition*. Elsevier.
- Wu, Xindong, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, et al. 2007. *Top 10 Algorithms in Data Mining. Knowledge and*

Information Systems. Vol. 14. doi:10.1007/s10115-007-0114-2.